外显生长指数☆

Ronald Rousseau [1,2,3]，胡小君[4,5]
1 KHBO (Association K.U.Leuven), Industrial Sciences and Technology, Zeedijk 101, B-8400 Oostende, Belgium
2 University of Antwerp, IBW, Venusstraat 35, B-2000 Antwerpen, Belgium
3 K.U.Leuven, Dept. Mathematics, Celestijnenlaan 200B, B-3000 Leuven (Heverlee), Belguim
4 浙江大学公共管理学院
5 浙江大学医学部医学信息中心
E-mail: ronald.rousseau@khbo.be，xjhu@zju.edu.cn

[摘　要] 基于一篇文章所引用的参考文献和该文的被引频次两个维度,提出了衡量该文外显成长度的相对指标。该研究涉及的引用数据及每条参考文献的被引数据均来自网络版科学引文索引(Web of Science)，检索日期为2010年4月最后一周。

[关键词] 引文分析　外显生长指数　参考文献　被引频次

1 引言

引文分析作为信息计量学的一个分支，涉及施引与被引、引用频次与引用模式的研究。这些研究表现为对作者、期刊、学科、以及其他相关单元的分析。此外，引文分析能从更深层次揭示被引与施引单元(如文献、作者、国家等等)之间的关系。从应用角度而言，引文分析可以看成是同行之间的协同努力，以分析并促进学术研究与论文质量[1,2]。

1960年代初，当加菲尔德开始从科技论文的参考文献和著录数据中收集信息时，他无法想象这将在研究评价领域引发一场风暴。如今，根据汤森路透及其竞争者的数据库所提供的引用数据而形成的科学计量学指标，以及对其用途与意义展开的探讨已经风靡学术界。*h*指数[3,4]、王冠指标[5,7]及英国研究评价实践的讨论[8,9]，便是这一现象的例证。也正因为如此，人们说加菲尔德引发了“蝴碟效应”[10,11]。

本文所要研究的问题是，根据一篇文章参考文献的被引用情况来描述和定义其成长程度。

为此，我们站在了加菲尔德毕生工作(实际上只是一部分，因为他的贡献远远超出这里所用到的)的基础上，利用了他的网络版科学引文索引(WoS)。

2 方法

我们主要关注JCR中分类为“普通内科学”的期刊上的文章，提取文章的被引次数，数据分为包含作者自引及剔除自引两种情况(数据取自WoS，提取时间为2010年4月最后一周)，同时提取每篇文章中每条参考文献的被引次数(同样来自WoS)。如果这些参考文献条目包含在WoS记录中，则直接记录下它的被引次数；否则，就利用“被引参考文献检索”功能，通过检索条目的匹配获取某一参考文献的被引次数。众所周知，由于拼写的变化或不完整的文献信息，很难获取精确的被引数据，但我们还是力图得到最合理、最接近的数值(书籍除外，因为书籍的被引次数比其他所有与外显生长指标计算相

☆该文经授权译自*Annals of Library and Information Studies*, 2010, 57:787-290

关的出版物高得多)。因此，我们也特别关注了参考文献条目中是否含有书籍。

我们认为文章发表10年后引文会趋向稳定。从文章发表年份开始，研究外显生长指数随时间变化的情况或许更有意思。这一点我们在文章结尾部分有所讨论。

外显生长指数的定义

如果文章*A*的参考文献列表中含有*R*条参考文献，我们将这些参考文献按各自的被引次数进行排列，然后将文章*A*按其被引次数也顺序加入到该列表中，则得到一个*R*+1项的列表。若被引次数相同，则取排列位次的平均数。如果文章*A*排在第R_A位，那么其指数就是：

$$I(A)=1-\frac{R_A}{R+1}$$

该指数取值在0(含)到1(不含)之间。此处，1不包含在内，因为排列不是从0开始，不能将*I*(A)定义为 $1-\frac{R_A}{R}$。如此选择是因为参考文献数量较少的文章比参考文献数量较多的文章更容易获得被引次数第一的位置。因此，如果某篇文章有9条参考文献，而它的被引次数超过这9条参考文献的被引次数，则它的外显生长指数是1 – 1/10 = 0.9；如果参考文献是49条，而文章的被引次数依然排在第一位，则它的外显生长指数是1 – 1/50 = 0.98。

除了该指数的通用定义，外显生长指数的计算还有其他一些变化，在下文的例子中可以看到。

(1) 剔除自引，参考文献列表变短(或不变)。

(2) 剔除书籍，因为一篇研究论文很难比一本标准参考书获得更多的引用。

(3) 剔除所有非WoS来源期刊的参考文献条目(如书等)，得到WoS外显生长指数。

另外，文章A的被引次数也可按含作者自引和不含自引两种方式计算。

3 结果

为便于说明，我们主要收集了一些高被引的医学文章，同时也有少量其他领域的文章。第一个例子就是科学计量学领域的一篇经典文章，即引入共引概念的Henry Small的文章。

标题：Cocitation in scientific literature - new measure of relationship between 2 documents

作者: Small H
来源: Journal of the American Society for Information Science，卷: 24, 期: 4，页码: 265-269, 出版年: 1973
被引次数: 474 (其中自引20 次)

该文章含有10条参考文献，他们的被引次数分别是(按降序排列，以分号隔开)：1 454; 448; 433; 273; 40; 30; 28; 27; 11; 5 次，因此其外显生长指数是1 – 2 /11 = 0.818。剔除自引后文章排名无影响(474变成454，仍位列引用次数第二)。 由于其中3条参考文献不是来自WoS期刊源(1本书，2篇会议文献)，剔除后 Henry Small的WoS外显生长指数是1 – 2/8 = 0.75。我们注意到Small在文中采取独立表格的形式对某些论文进行了分析，WoS 似乎把这些论文都当成正常参考文献而囊括其中。如果按WoS给出的参考文献列表则有18条参考文献，这些加进去的参考文献都有很高的被引次数，这样Small的文章根据被引次数排名就变成了第10位。以这种方式计算，Small的外显生长指数只有1 – 10/19 = 0.474。

3.1 医学领域高被引文章外显生长指数计算举例

例1：
标题: Statistical-methods for assessing agreement between 2 methods of clinical measurement

作者: Bland JM, Altman DG
来源: Lancet，卷：327，期: 8476，页码: 307-310, 出版年：1986
被引次数: 16 906 (剔除自引后16 868)

这篇文章含8条参考文献，被引次数分别为：1 195; 130; 41; 21; 12; 3; 3; 1次。因此，该文的外显生长指数是1 – 1/9 = 0.889。剔除自引后外显生长指数无差别。由于其中的4条参考文献不是WoS期刊源，故该文的WoS外显生长指数是1 – 1/5 = 0.8。

例2：
标题: Assessment of outcome after severe brain damage practical scale

作者: Jennett B, Bond M
来源: Lancet，卷: 305，期: 7905，页码: 480-484, 出

版年: 1975
被引次数: 3 259 (剔除自引后3 216)

这篇文章含24条参考文献，被引次数分别为：4 350; 460; 151; 112; 105; 96; 89; 77; 72; 67; 65; 59; 53; 29; 27; 23; 12; 6; 5; 4; 2; 1; 1; 1 次。因此，该文的外显生长指数是1 – 2/25 = 0.92，剔除自引后无差异。由于其中13条文献(不含被引次数最多的那条)不是WoS期刊源，故该文的WoS外显生长指数是1 – 2/12 = 0.833。

例3：
标题: Unidentified curved bacilli in the stomach of patients with gastritis and peptic-ulceration

作者: Marshall BJ, Warren JR
来源: Lancet，卷: 323，期：8390，页码: 1311-1315, 出版年: 1984
被引次数: 2 359 (其中自引1次)

这篇文章含29条参考文献，被引次数分别是：3 621; 2 594; 1 395; 753; 565; 557; 557; 458; 388; 326; 319; 239; 226; 158; 142; 136; 121; 110; 102; 94; 64; 52; 38; 30; 29; 16; 6; 2; 1 次。因此该文的外显生长指数是1 – 3/30 = 0.9，剔除自引后无差异。由于其中8条文献(不含被引次数最多的2条)不是WoS期刊源，故该文的WoS外显生长指数是1 – 3/22 = 0.864。

例4：
标题: A new simplified acute physiology score (saps-ii) based on a European North-American multicenter study

作者: Legall JR, Lemeshow S, Saulnier F
来源: JAMA-Journal of the American Medical Association，卷: 270，期: 24，页码: 2957-2963，出版年: 1993
被引次数: 1 805 (其中自引3次)

这篇文章含21条参考文献，被引次数分别是：5 849; 5 791; 2 732; 1 555; 914; 841; 732; 698; 558; 449; 221; 202; 185; 156; 114; 82; 66; 62; 43; 7; 3次。因此该文的外显生长指数是1 – 4/22 = 0.818，剔除自引后无差异。由于其中5条参考文献(不含被引次数最多的3条)不是WoS期刊源，故该文的WoS外显生长指数是1 – 4/17 = 0.765。

3.2 参考文献数量少的文章的外显生长指数计算举例

标题: Prematurity and uniqueness in scientific discovery

作者: Stent GS
来源: Scientific American，卷: 227，期: 6，页码: 84-90，出版年: 1972,
被引次数: 120 (其中自引1次)

这篇文章含3条参考文献：

[1] Churchman, C.W. Hansel CEM – ESP – A scientific evaluation. Science, 1966, 153 (3740), pp.1088-. (被引3 次).

[2] Polanyi, M. Potential theory of adsorption. Science, 1963, 141(358), pp. 1010-. (被引50次)

[3] Stent, G.S. Molecular Genetics. An introductory narrative. San Francisco: W.H. Freeman, 1971.(这是一本书，因而不是WoS期刊源，被引60次)

这篇文章的被引次数(包含或剔除自引)大于它的3篇参考文献，其外显生长指数的最大值是1 – 1/4 = 0.75。剔除书、或剔除非WoS来源条目(在此都一样)，外显生长指数降低到0.667。这个例子说明在一个短参考文献列表上外生相对容易。该例子中，参考文献列表中含有一本被引次数相当高的书，但是并没有高过我们进行外显生长指数测定的这篇文章。

3.3 外显生长指数为零的例子

最后列举一篇在参考文献列表上根本没有外生的文章(我们自己的一篇文章)。

标题: A characterization of distributions which satisfy price law and consequences for the laws of zipf and mandelbrot

作者: Egghe I, Rousseau R
来源: Journal of Information Science, 卷: 12，期: 4，页码: 193-197, 出版年: 1986
被引次数: 6 (其中2篇为Egghe自引)

这篇文章含9条参考文献，然而其中2条是未发表的(在那个时代)，不能计算在内。余下7条的被引次数排列是：(> 600); 328; 45; 28; 18; 17; 1次，因此该文的外显生长指数是1 – 7/8 = 0.125。 这其中又有两条参考文献：被引最多的(一本书)和被引最少的(百科全书中的一个条目)不是WoS来源项，因而该文

的WoS外显生长指数是1 – 6/6 = 0。剔除自引后结果无差异。

研究WoS出版物的参考文献时，我们注意到所有参考文献条目都至少有1次被引。有趣的是，在很多样例中，许多参考文献都恰恰只被引用过1次。

4 结论与评论

本文的撰写遵循了信息科学家金•加菲而德的探索精神。这不是一篇成熟的科学论文，而是希望在引文数据的基础上提出一个有意思的指标。

外显生长指数与Ortega假说有某种关联，但我们的目标更加适度，我们只是想找到一种合理的描述和表示方法。外显生长指数也许可以用来检验Ortega假说。Ortega假说阐述的是，科学精英的工作在很大程度上要归功于普通科学家的工作，伟大科学家的工作则是在这些普通科学家的小发现的基础上搭建金字塔[12-15]。

根据这些例子的研究，计算或不计算自引似乎对结果的影响不大。毫无疑问，高被引文章容易在参考文献基础上外生。当然，要想得出一般性结论，还需要对大量的外显生长指数进行测定。

按照某种可能的时间序列[16]，研究外显生长指数随时间的变化，会是下一步要做的有意思的事。

参考文献

[1] Spinak E. Diccionario Enciclopédico de Bibliometría, Cienciometría e Informetría. Caracas: UNESCO-CII/II, 1966.

[2] Rousseau R. Publication and citation analysis as a tool for information retrieval. In: Social Information Retrieval Systems. Emerging technologies and applications for searching the web effectively (Goh D & Foo S, eds.). Hershey (PA); Information Science Reference (IGI Global). 2008, Chapter 13, pp. 252-267.

[3] Hirsch J E. An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102: 16569-16572.

[4] Waltman L, Van Eck N J. A simple alternative to the h-index. ISSI Newsletter, 2009, 5(3):46–48.

[5] Lundberg J. Lifting the crown—citation z-score. Journal of Informetrics, 2007, 1(2):145–154.

[6] Opthof T, Leydesdorff L. Caveats for the journal and field normalizations in the CWTS (“Leiden”) evaluations of research performance. Journal of Informetrics, 2010, 4(3):423-430.

[7] Waltman L, Van Eck N J, Van Leeuwen T N, Visser M S, Van Raan A F J. Towards a new crown indicator: Some theoretical considerations. 2010, arXiv:1003,2167.

[8] Warner J. A critical review of the application of citation studies to the Research Assessment Exercises. Journal of Information Science, 200, 026(6):453-459.

[9] Oppenheim C. Out with the old and in with the new: the RAE, bibliometrics and the new REF. Journal of Librarianship and Information Science, 2008, 40(3):147-149.

[10] Lorenz E N. Deterministic nonperiodic flow. Journal of the Atmospheric Sciences, 1963, 20(2):130-141.

[11] Gleick J. Chaos: Making a new science. New York: Viking, 1987.

[12] Cole J R. The social structure of science: a study of the reward and communications systems of modern physics. Ph.D. dissertation: Columbia University, 1969.

[13] Cole J R, Cole S. The Ortega hypothesis: citation analysis suggests that only a few scientists contribute to scientific progress. Science, 1972, 178: 368-375.

[14] Hoerman H L, Nowicke C E. Secondary and tertiary citing: a study of referencing behavior in the literature of citation analysis deriving from the Ortega hypothesis of Cole and Cole. Library Quarterly, 1995, 65(4):415-434.

[15] Rousseau R, Yang LY, Yue T. A discussion of Prathap's h_2-index for institutional evaluation with an application in the field of HIV infection and therapy. Journal of Informetrics, 2010, 4(2):175-184.

[16] Liu YX, Rousseau R. Definitions of time series in citation analysis with special attention to the h-index. Journal of Informetrics, 2008, 2(3):202-210.

An Outgrow Index

Ronald Rousseau [1,2,3], *Hu Xiaojun* [4,5]

1 KHBO (Association K.U.Leuven), Industrial Sciences and Technology, Zeedijk 101, B-8400 Oostende, Belgium
2 University of Antwerp, IBW, Venusstraat 35, B-2000 Antwerpen, Belgium
3 K.U.Leuven, Dept. Mathematics, Celestijnenlaan 200B, B-3000 Leuven (Heverlee), Belguim
4 College of Public Administration, Zhejiang University, Hangzhou, China
5 Medical Information Centre, Zhejiang University School of Medicine, Hangzhou, China
E-mail: ronald.rousseau@khbo.be, xjhu@zju.edu.cn

[Abstract] Proposes a relative index measuring the amount by which an article outgrows, in terms of citations, the publications on which it is based. The study involves citations collected from Web of Science during the last week of April 2010 along with the number of citations received (also in WoS) by each of the references.

[Keywords] citation analysis, outgrow index, references, citations

科学新闻

基因组复杂性进化的能量学

所有复杂生命体都是由真核细胞构成的。真核细胞从原核细胞进化而来在40亿年漫长历史过程中仅发生了一次，否则原核细胞将不能表现出演化出更高复杂性的趋势。这是因为，原核细胞基因组的规模受到生物能量学的限制。英国伦敦大学学院Nick Lane和德国Düsseldorf 大学William Martin研究推测，相对于提供生物能量的质膜，因胞内共生而导致的线粒体的进化形成重新建构了DNA的分布格局，它使得基因数目发生显著的约20万倍的扩增。基因组规模的这种极大的跳跃严格依赖于线粒体所能提供的能量，这是真核细胞复杂性进化形成的前提，也是通向多细胞生物的关键。相关研究论文发表在2010年10月21日*Nature*[467(7318):929-934]上。

将人成纤维细胞直接转化为造血前体细胞

利用重编程得到的人诱导性多能干细胞分化为所需要的细胞受制于人们对世系特异性的了解程度。加拿大McMaster 大学Mickie Bhatia与合作者，演示了不需要建立多能性，而是将真皮成纤维细胞直接转化为造血前体细胞和成熟细胞的可行性。异位表达OCT4激活的造血转录因子，辅以特异的细胞因子，将导致表达泛白细胞标记CD45的细胞的生成。这些源于成纤维细胞的独特细胞植入体内后可分化产生粒细胞、单核细胞、巨核细胞和红系细胞。该研究显示，绕过形成多能性的过程可直接将成纤维细胞转化为造血前体细胞，这提出了一种替代细胞重编成获得自体细胞用于细胞移植治疗的方法，它可以避免采用重编成人多能性干细胞具有的多种安全隐患。相关研究论文发表在2010年11月25日*Nature*[468(7323):521-526]上。

31个大豆基因组重测序揭示遗传多样性和进化选择模式

香港中文大学农业生物技术国家重点实验室Samuel Sai-Ming Sun和Hon-Ming Lam研究组与深圳华大基因研究院Gengyun Zhang和王俊及其他合作者，对大豆全基因组变异模式进行了大规模分析。研究人员对17个野生和14个栽培大豆品种的基因组进行了重测序，测序深度达到5倍，覆盖率达到90%以上。研究人员对比了野生品种和栽培品种的遗传变异模式，结果发现野生品种具有更高的等位遗传多样性。他们在大豆基因组中识别出高水平的连锁不平衡，这提示大豆的标记辅助育种相比基因图位克隆不具有优势。研究人员获得了连锁不平衡区块的位置和分布，发现了205 614个单核苷多态性位点。这些位点将有助于数量性状位点定位和关联分析。该研究为野生大豆研究、未来育种以及数量性状分析提供了有价值的资源。相关研究论文发表在2010年12月*Nature Genetics*[42(12):1053-1059]上。